Data Analytics Project

ABSENTEEISM (During working hours)

CASE STUDY

Data Analytics Project

Submitted By :

Roshan Jaiswal (2021BCS-079)

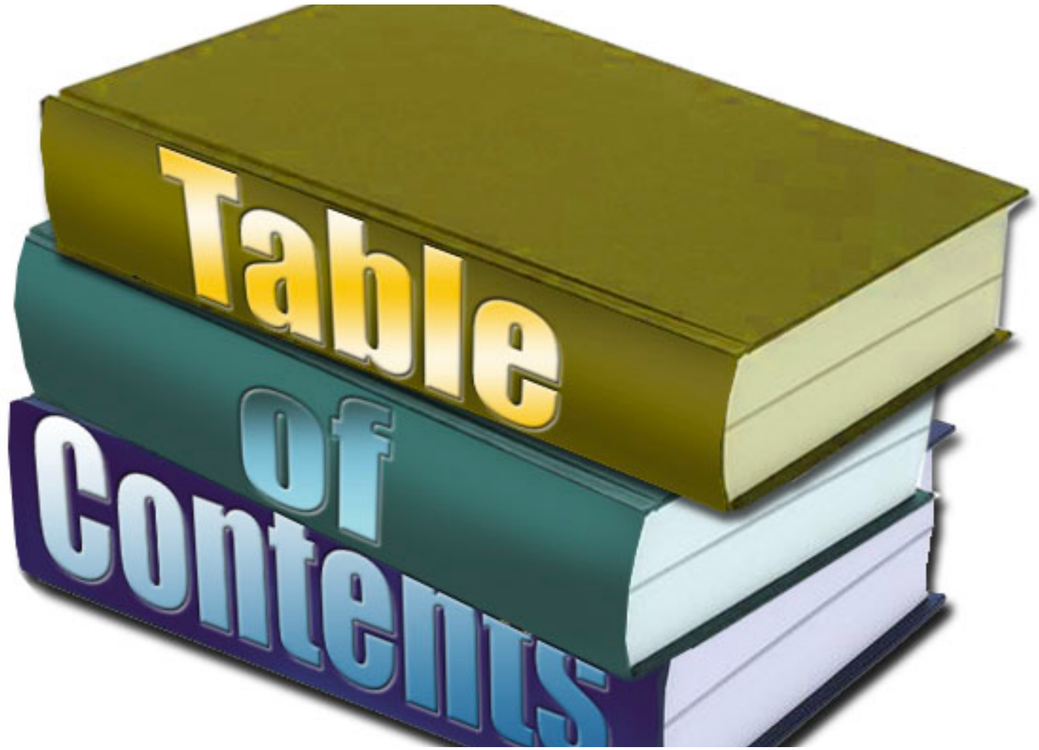
Vedant Maske (2021BCS-069)

Gagandeep Singh (2021BCS-026)

Submitted to :

Prof. Santosh Singh Rathore

TABLE OF CONTENT



INTRODUCTION

- Absenteeism includes situations where employees are frequently absent.
- This absence from work during working hours results in temporary incompetence in executing regular working activity in the company.

DATASET DURATION

POINT OF VIEW OF ANALYSIS

AIM

- Our primary goal is to determine the duration of an employee's absence from work in terms of working hours.

REASON

- Highr competitiveness
- Unachievable business goals
- Elevated risk of becoming unemployed

RISK

- Absenteeism harms productivity and endangers morale, team dynamics, and overall performance.

DATASET

	ID	Reason for Absence	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours
0	11	26	07/07/2015	289	36	33	239.554	30	1	2	1	4
1	36	0	14/07/2015	118	13	50	239.554	31	1	1	0	0
2	3	23	15/07/2015	179	51	38	239.554	31	1	0	0	2
3	7	7	16/07/2015	279	5	39	239.554	24	1	2	0	4
4	11	23	23/07/2015	289	36	33	239.554	30	1	2	1	2
5	3	23	10/07/2015	179	51	38	239.554	31	1	0	0	2
693	25	10	21/05/2018	235	16	32	237.656	25	3	0	0	8
694	15	22	23/05/2018	291	31	40	237.656	25	1	1	1	8
695	17	10	23/05/2018	179	22	40	237.656	22	2	2	0	8
696	28	6	23/05/2018	225	26	28	237.656	24	1	1	2	3
697	18	10	24/05/2018	330	16	28	237.656	25	2	0	0	8

FEATURES

01

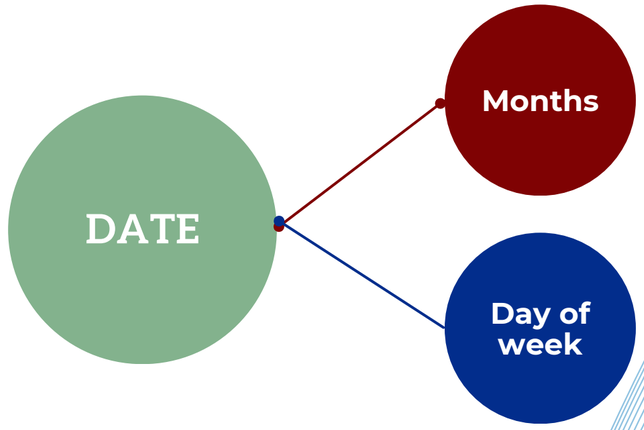
ID: Nominal

Employee identification number.

02

Date: string -> ordinal

The date of the absence.



03

Transportation Expense:Numerical

The amount spent by the employee on transportation.

FEATURES

04

Distance to Work:Numerical

The distance in kilometers between the employee's home and workplace.

05

Age: Discrete (Numerical)

The age of the employee.

06

Daily Work Load Average:Numerical

The average amount of work an employee handles in a day.

07

BMI: Continuous (Numerical)

A measure of body fat based on an individual's weight and height.

FEATURES

08

Pets: Discrete (Numerical)

The number of pets the employee has.

09

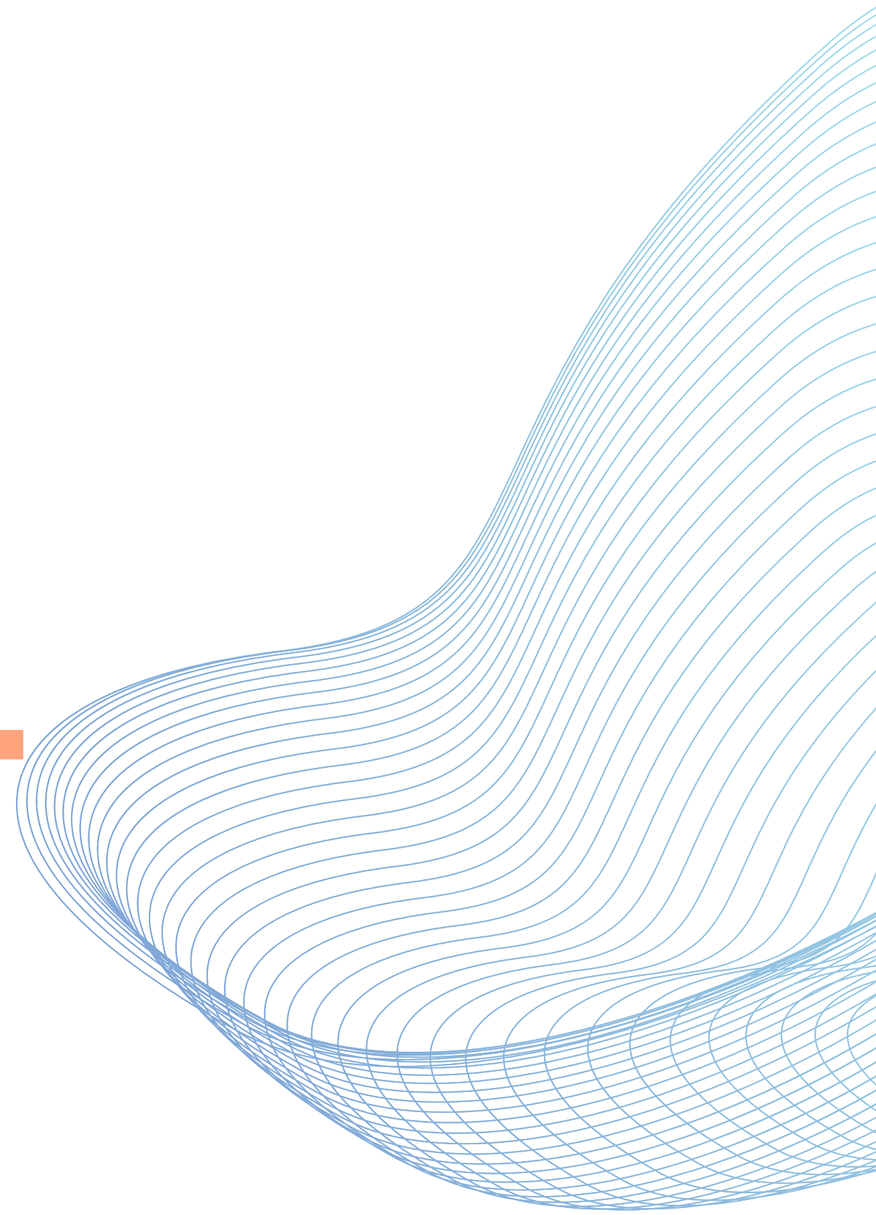
Children:Discrete (Numerical)

The number of children the employee has.

10

Absenteeism Time in Hours:Continuous (Numerical)

The total hours of absenteeism for the employee.



11 Reason for Absence: Nominal

1. Certain infectious and parasitic diseases
2. Neoplasms
3. Diseases of the blood and blood-forming organs, and disorders involving the immune mechanism
4. Endocrine, nutritional, and metabolic diseases
5. Mental and behavioral disorders
6. Diseases of the nervous system
7. Diseases of the eye and adnexa
8. Diseases of the ear and mastoid process
9. Diseases of the circulatory system
10. Diseases of the respiratory system
11. Diseases of the digestive system
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Diseases of the genitourinary system
15. Pregnancy, childbirth, and the puerperium
16. Certain conditions originating in the perinatal period
17. Congenital malformations, deformations, and chromosomal abnormalities
18. Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified
19. Injury, poisoning, and certain other consequences of external causes
20. External causes of morbidity and mortality
21. Factors influencing health status and contact with health services.

11

Reason for Absence: Nominal

22. Patient follow-up
23. Medical consultation
24. Blood donation
25. Laboratory examination
26. Unjustified absence
27. Physiotherapy
28. Dental consultation

- **Removal of multicollinearity**
- **Example**
- **Dimensionality Reduction**

We Classified 28 reasons into 4 groups for simplicity and improved analysis.

In [9]:
```
reason_columns = reason_columns.drop(['check'], axis = 1)


reason_columns = pd.get_dummies(df['Reason for Absence'], drop_first = True)
reason_columns
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	21	22	23	24	25	26	27
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False
3	False	False	False	False	False	False	True	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False
5	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False
6	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False	False
7	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False
8	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False	False
10	True	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False

CLASSIFICATION

1 **GROUP A**
FROM 1 TO 14

....
. Endocrine, nutritional, and metabolic diseases
. Mental and behavioral disorders
. Diseases of the nervous system
. Diseases of the eye and adnexa
....

2 **GROUP B**
FROM 15 TO 17

15. Pregnancy, childbirth, and the puerperium
16. Certain conditions originating in the perinatal period
17. Congenital malformations, deformations, and chromosomal abnormalities

3 **GROUP C**
FROM 18 TO 21

18. Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified
19. Injury, poisoning, and certain other consequences of external causes
20. External causes of morbidity and mortality
21. Factors influencing health status and contact with health services.

4 **GROUP D**
FROM 22 TO 28

22. Patient follow-up
23. Medical consultation
24. Blood donation
25. Laboratory examination
26. Unjustified absence
27. Physiotherapy
28. Dental consultation

In [10]:
```
df = df.drop(['Reason for Absence'], axis = 1)



reason_type_1 = reason_columns.loc[:, 1:14].max(axis=1)
reason_type_2 = reason_columns.loc[:, 15:17].max(axis=1)
reason_type_3 = reason_columns.loc[:, 18:21].max(axis=1)
reason_type_4 = reason_columns.loc[:, 22:].max(axis=1)
```

In [11]:
```
df = pd.concat([df, reason_type_1, reason_type_2, reason_type_3, reason_type_4], axis = 1)
df
```

Out[11]:

	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours	0	1	2	3
0	07/07/2015	289	36	33	239.554	30	1	2	1	4	False	False	False	True
1	14/07/2015	118	13	50	239.554	31	1	1	0	0	False	False	False	False
2	15/07/2015	179	51	38	239.554	31	1	0	0	2	False	False	False	True
3	16/07/2015	279	5	39	239.554	24	1	2	0	4	True	False	False	False
4	23/07/2015	289	36	33	239.554	30	1	2	1	2	False	False	False	True
5	10/07/2015	179	51	38	239.554	31	1	0	0	2	False	False	False	True
6	17/07/2015	361	52	28	239.554	27	1	1	4	8	False	False	False	True

12 Education: Ordinal ->Binary

Group 1:- 0
Group 2:- 1

GROUP 1

H

HIGH SCHOOL

LEVEL 1

G

GRADUATE

LEVEL 2

GROUP 2

P

POSTGRADUATE

LEVEL 3

M

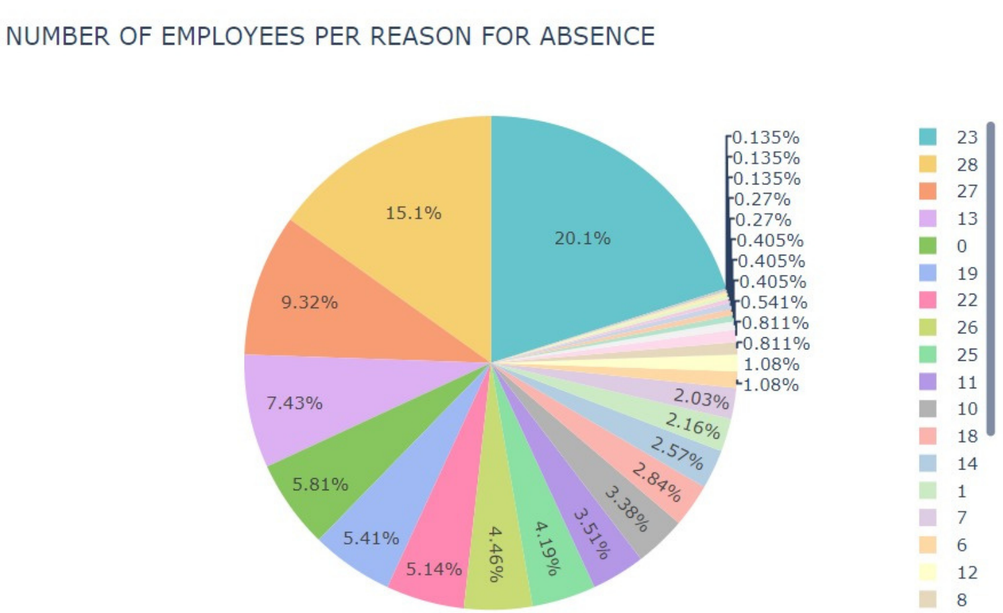
MASTER OR DOCTOR

LEVEL 4

Data After Preprocessing

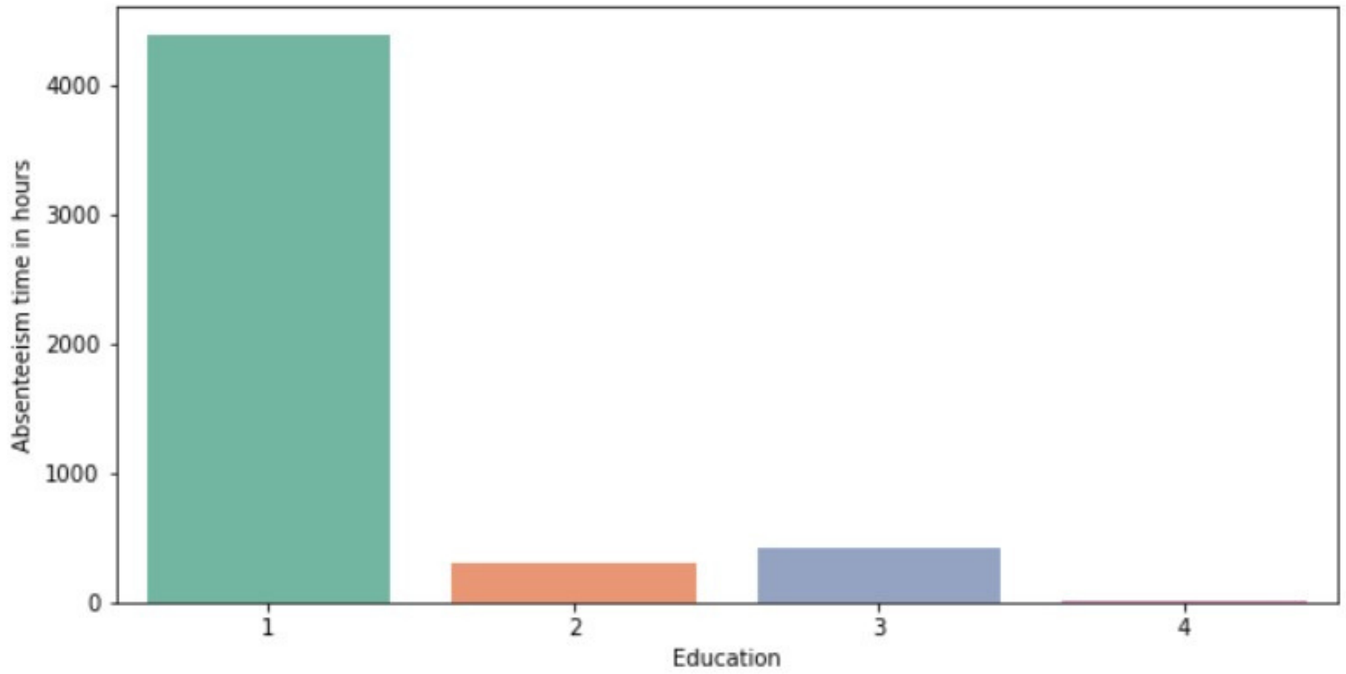
	Reason_1	Reason_2	Reason_3	Reason_4	Month Value	Day of the Week	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours
0	False	False	False	True	7	1	289	36	33	239.554	30	0	2	1	4
1	False	False	False	False	7	1	118	13	50	239.554	31	0	1	0	0
2	False	False	False	True	7	2	179	51	38	239.554	31	0	0	0	2
3	True	False	False	False	7	3	279	5	39	239.554	24	0	2	0	4
4	False	False	False	True	7	3	289	36	33	239.554	30	0	2	1	2
5	False	False	False	True	7	4	179	51	38	239.554	31	0	0	0	2
6	False	False	False	True	7	4	361	52	28	239.554	27	0	1	4	8
7	False	False	False	True	7	4	260	50	36	239.554	23	0	4	0	4

Anlaysis and Finding



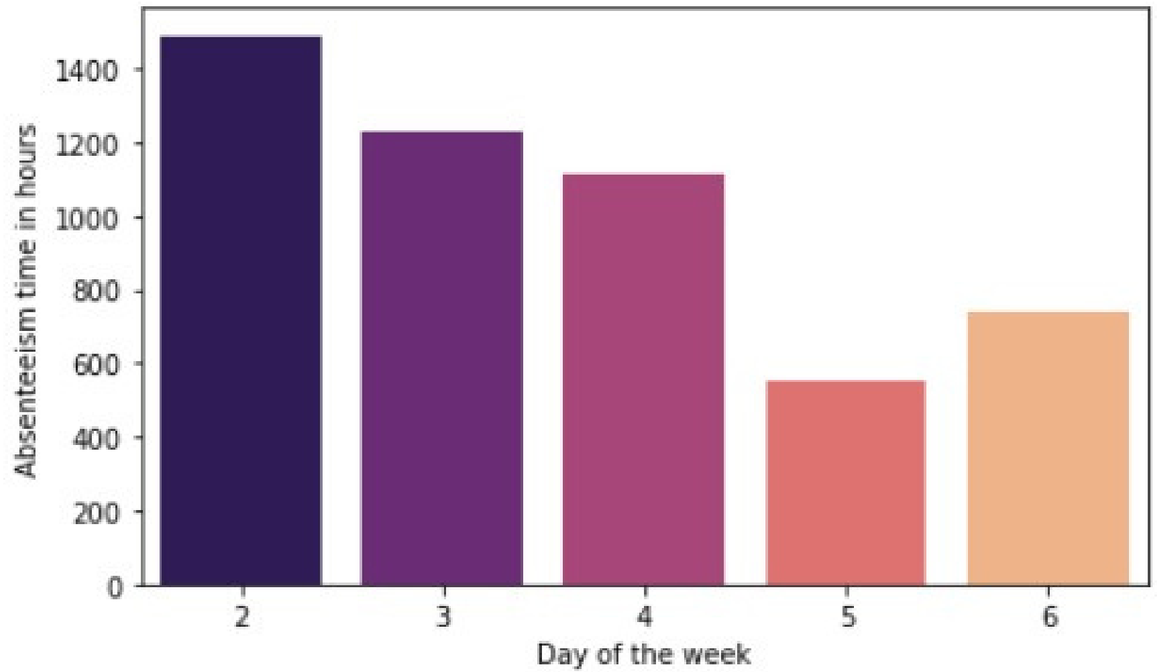
Inference

This pie chart displays the distribution of employee absences based on specific reasons. The most prevalent cause is medical consultation (23), constituting 20.1%, followed by dental consultation (28) at 15.1%, and physiotherapy (27) at 9.32%. These three reasons collectively contribute to over **45%** of all absences. To address this, managers should consider measures such as having an on-site clinical doctor to provide immediate medical support and reduce absenteeism, ultimately enhancing productivity.



Inference

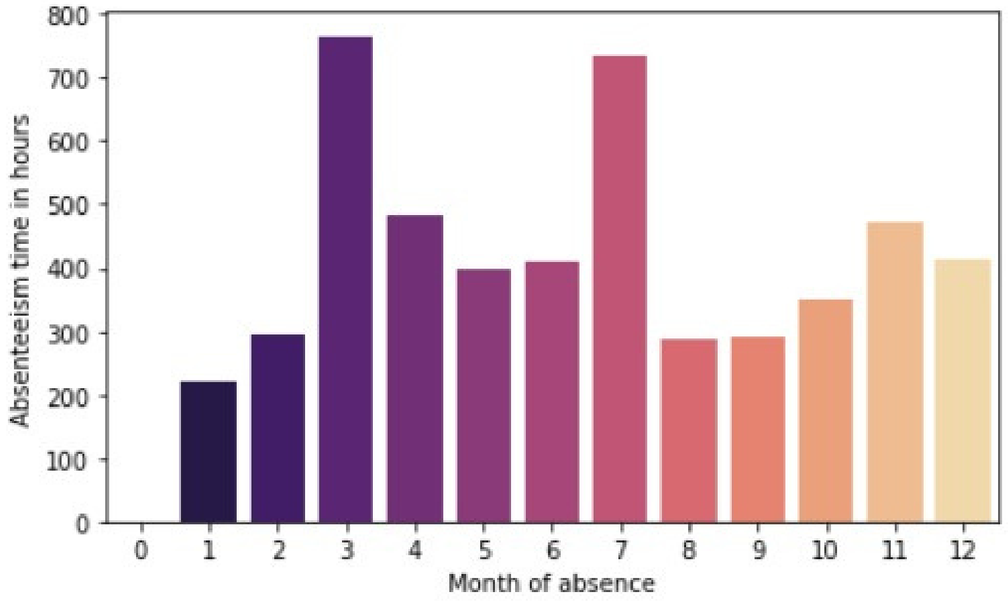
The bar graph illustrates absenteeism hours based on employees' education levels. Education level 1 records the highest hours (4500), followed by level 3 (500), level 2 (400), and level 4 (50). This indicates a correlation between lower education levels and increased absenteeism. Employers should address this by implementing targeted training programs for employees with lower education, fostering skill development and enhancing job performance.



Inference

This bar graph depicts the correlation between absenteeism hours and the workdays of the week, where 1 represents Sunday and 7 represents Saturday. Notably, absenteeism peaks on day 2 (Monday) and reaches its lowest on day 5. This pattern suggests a tendency for higher absenteeism at the beginning of the week, possibly influenced by post-weekend factors like hangovers or extended breaks. Conversely, lower absenteeism on Thursdays implies employees strategically completing tasks on that day to enjoy an extended weekend.

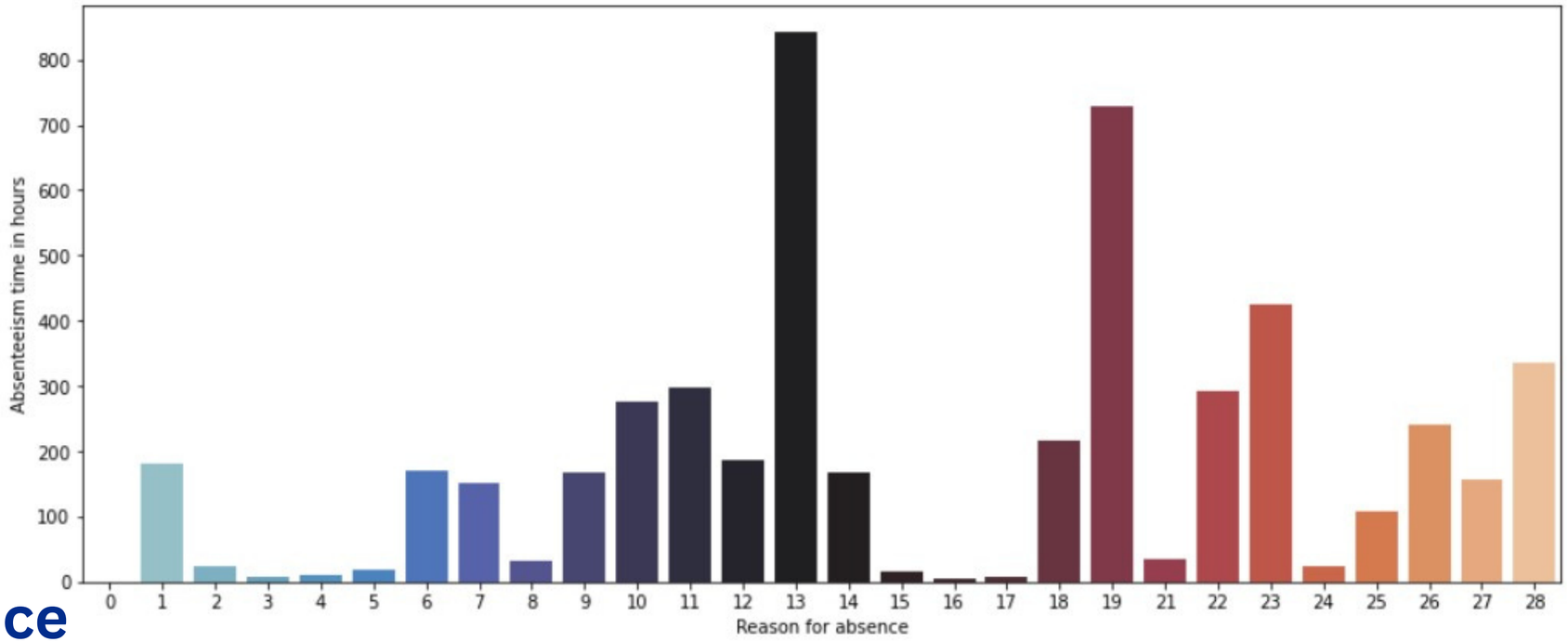
To address absenteeism, the manager can introduce flexible scheduling.



Inference

This graph shows the relationship between absenteeism in hours and the 12 months of the year. The graph shows that absenteeism is highest in the months of March and July, and lowest in the months of January.

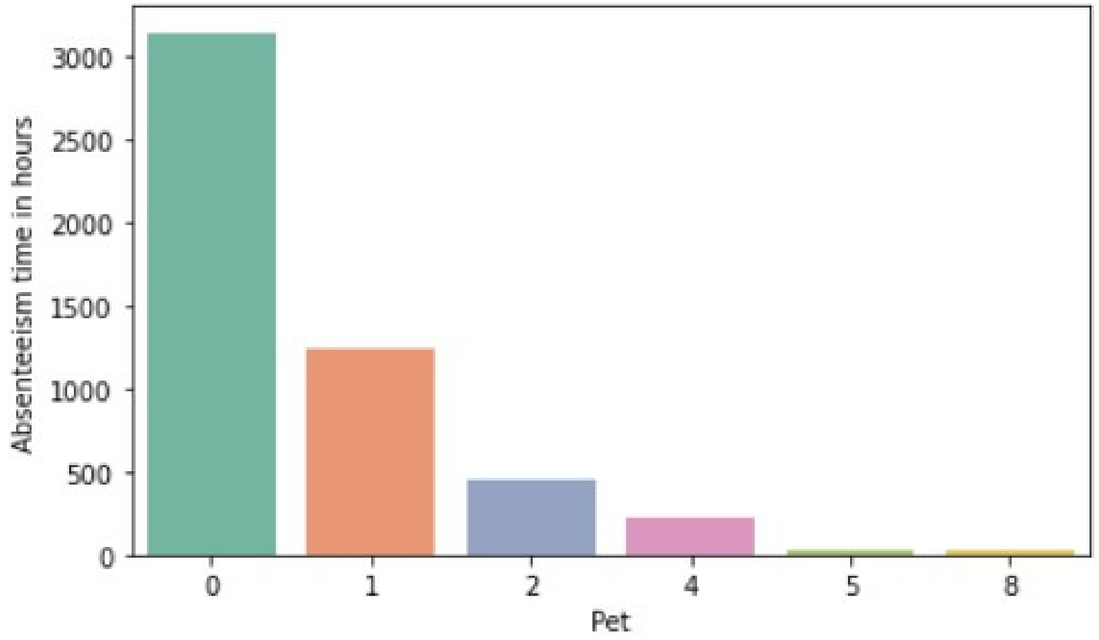
This could be due to the fact that people are more likely to take vacations during the early summer/spring monsoon months, and less likely to take time off during the winter months. It is also possible that the holiday season in December and January may lead to lower absenteeism rates.



Inference

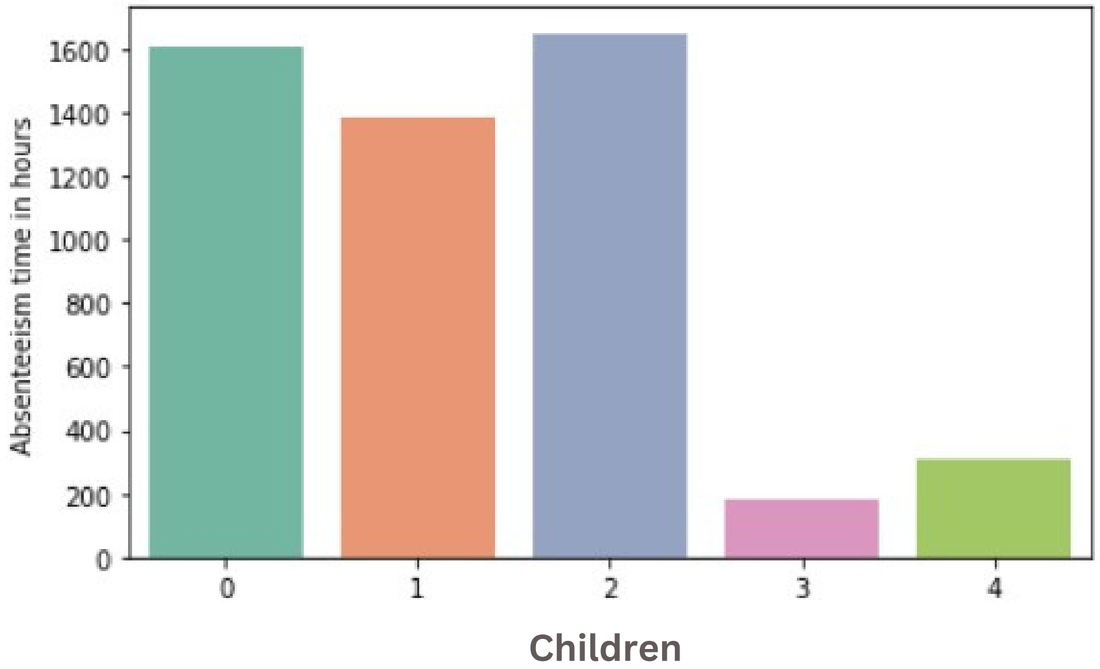
This bar graph illustrates the correlation between absenteeism hours and the reason for absence, ranging from 1 (Certain infectious and parasitic diseases) to 28 (Dental consultation). Notably, absenteeism peaks at 13 (Diseases of the musculoskeletal system and connective tissue) and reaches its lowest at 16 (Certain conditions originating in the perinatal period), followed by 19 (Injury, poisoning, and certain other consequences of external causes). This pattern indicates that the highest number of absenteeism hours is associated with reason 13, followed by reason 19.

To address absenteeism, the manager can implement targeted health and wellness initiatives, providing support for employees dealing with musculoskeletal issues and injuries.



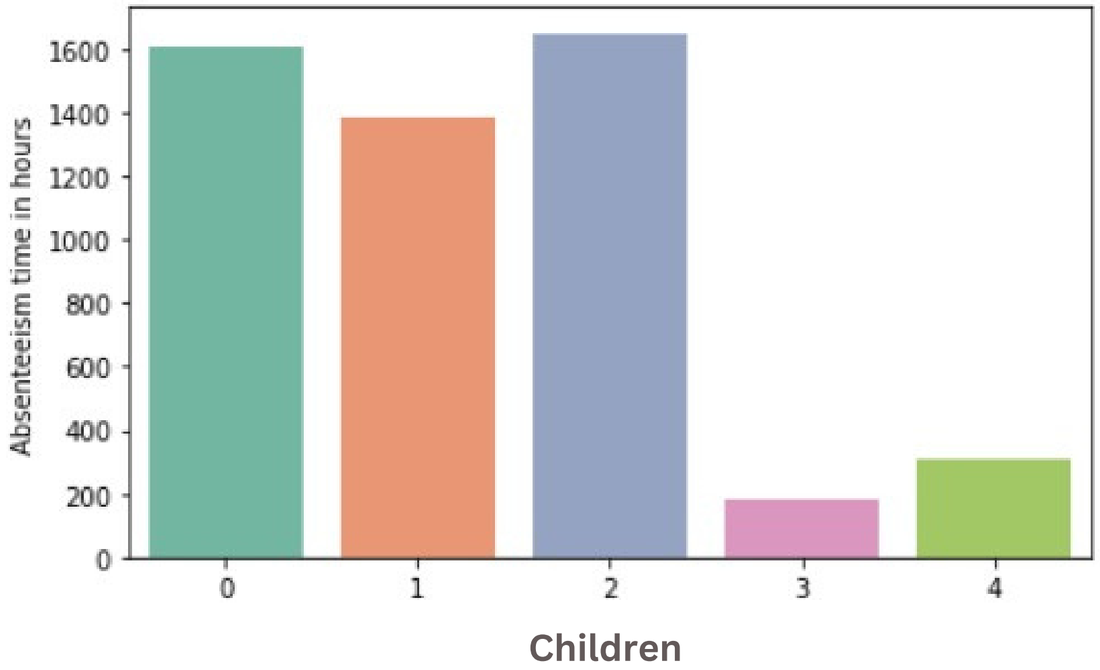
Inference

This bar graph illustrates the correlation between absenteeism hours and the number of pets, ranging from 0 (no pets) to 8 (8 pets). Interestingly, absenteeism peaks when employees have 0 pets and decreases as the number of pets increases, reaching its lowest point at 8 pets. This pattern may indicate higher absenteeism among individuals without pets, influenced by factors such as *personal health concerns or family responsibilities*. On the other hand, lower absenteeism at 8 pets suggests employees prioritize attendance, possibly by arranging for someone to supervise their pets. To address this, the manager can consider introducing onsite pet care services, particularly for employees with fewer than 2 pets, promoting a pet-friendly workplace and reducing absenteeism.



Inference

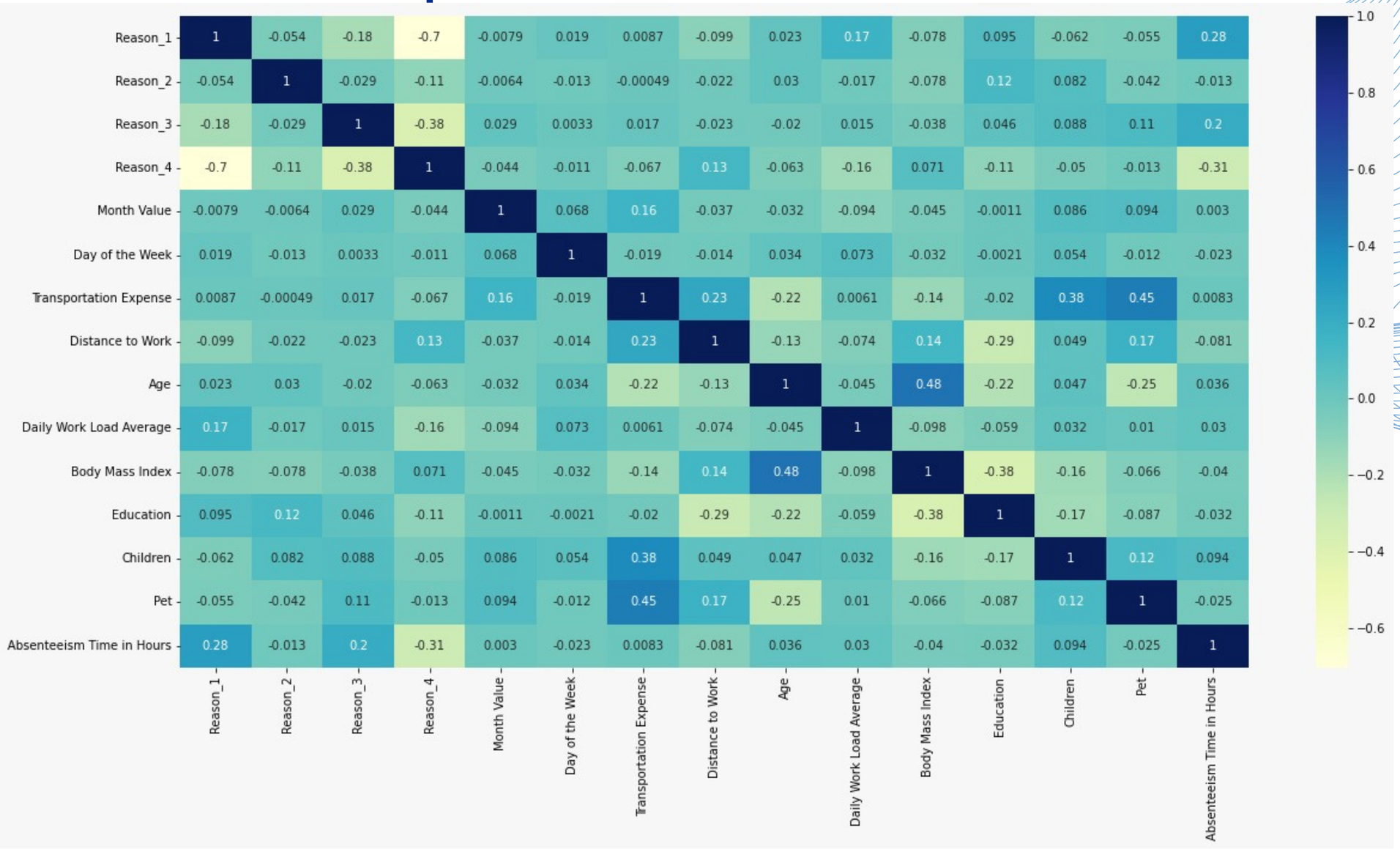
This bar graph outlines the connection between absenteeism hours and the number of children, ranging from 0 (no children) to 4 (4 children). Notably, absenteeism peaks when employees have 2 children, while it is lowest for those with 3 and 4 children. This trend suggests that individuals without children may experience higher absenteeism, and those with 1 or 2 children are influenced by factors such as childcare responsibilities. Conversely, lower absenteeism for employees with 3 or 4 children indicates a prioritization of attendance, possibly facilitated by arranging supervision for their children.
To address this, the manager can explore introducing onsite childcare facilities



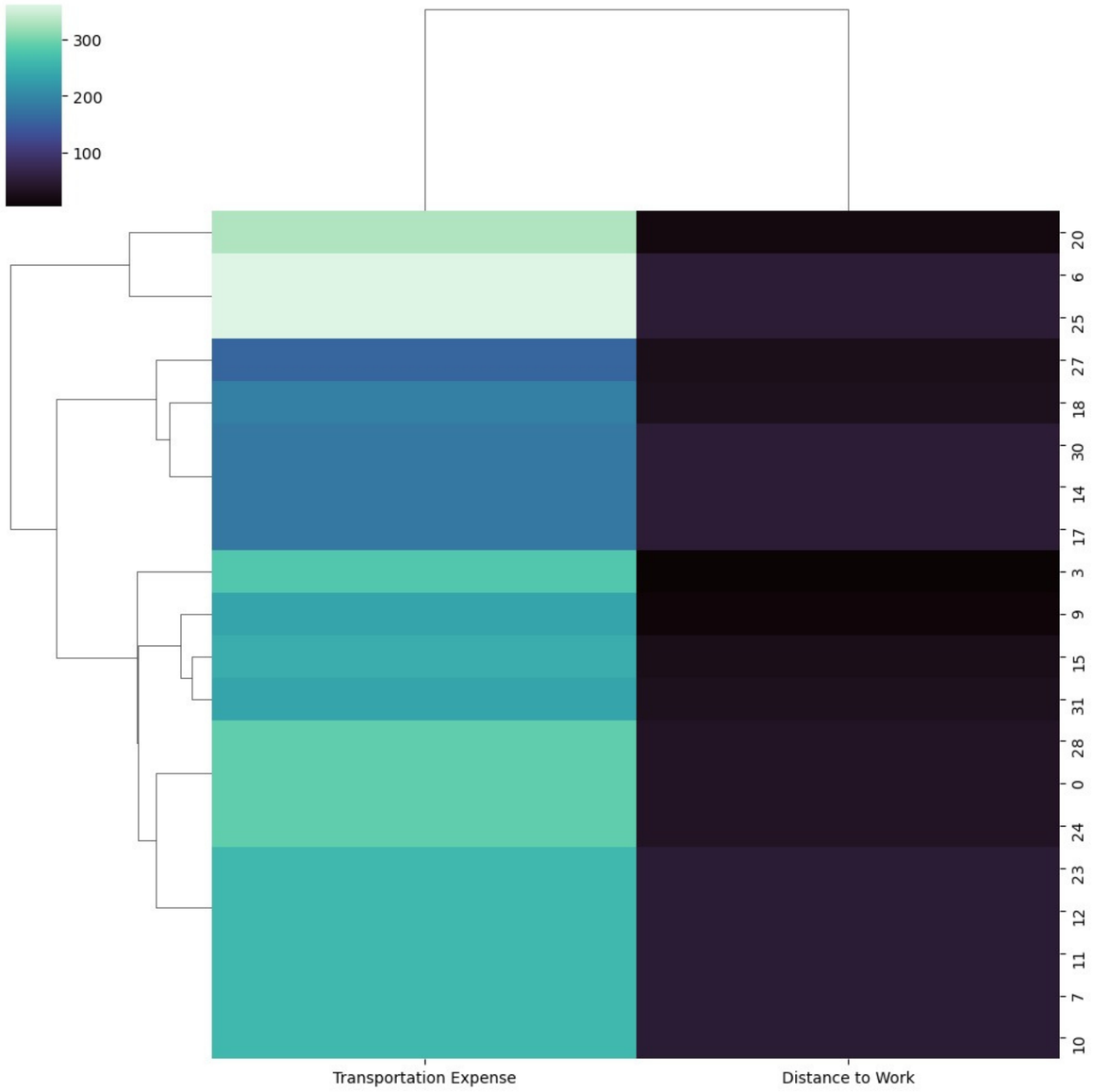
Inference

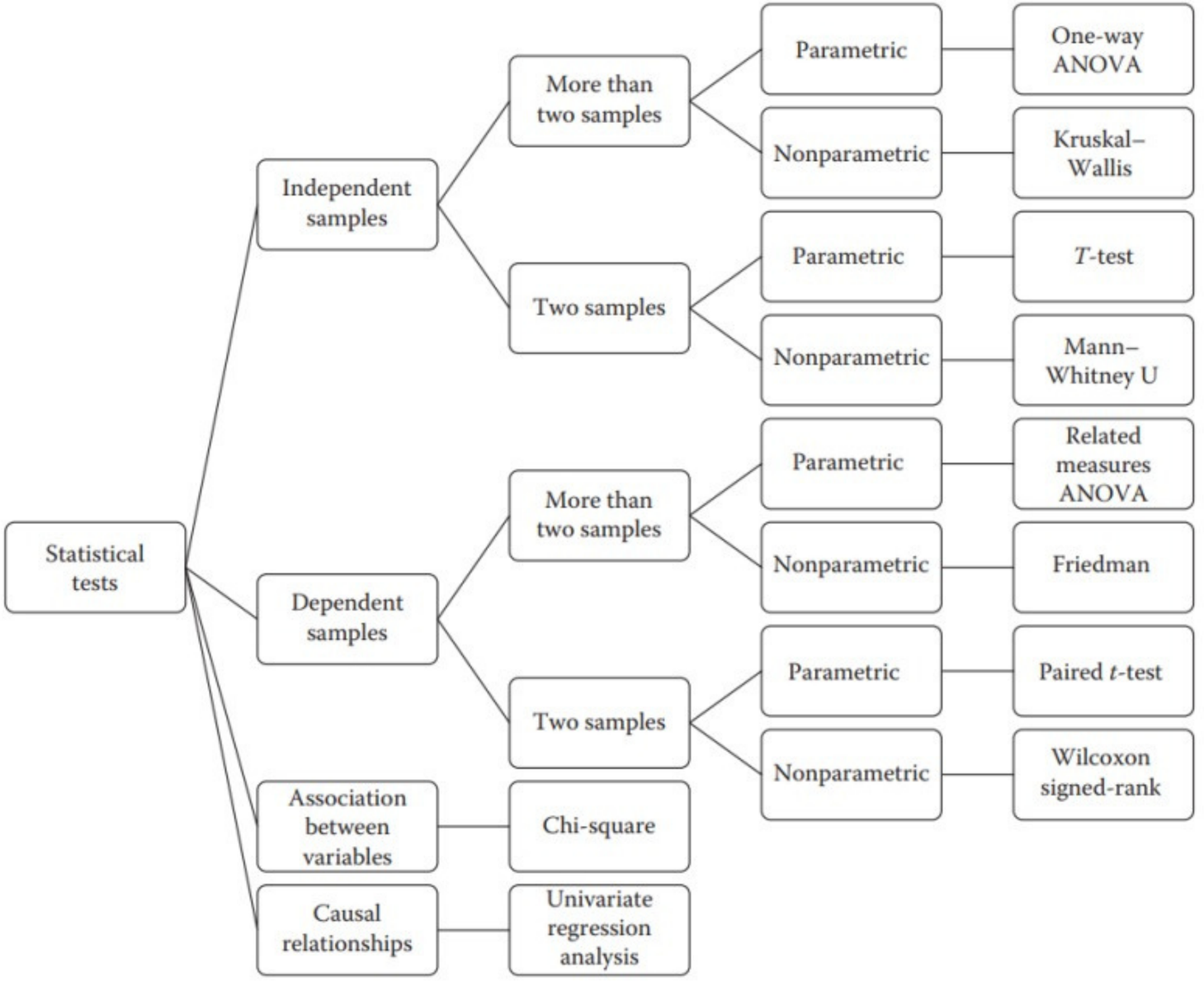
This bar graph outlines the connection between absenteeism hours and the number of children, ranging from 0 (no children) to 4 (4 children). Notably, absenteeism peaks when employees have 2 children, while it is lowest for those with 3 and 4 children. This trend suggests that individuals without children may experience higher absenteeism, and those with 1 or 2 children are influenced by factors such as childcare responsibilities. Conversely, lower absenteeism for employees with 3 or 4 children indicates a prioritization of attendance, possibly facilitated by arranging supervision for their children.
To address this, the manager can explore introducing onsite childcare facilities

Correlation heatmap



Heatmap and Dendogram




Statistical tests
Independent samples
More than two samples
Parametric
One-way ANOVA
Nonparametric
Kruskal–Wallis
Two samples
Parametric
T-test
Nonparametric
Mann–Whitney U
Dependent samples
More than two samples
Parametric
Related measures ANOVA
Nonparametric
Friedman
Two samples
Parametric
Paired t-test
Nonparametric
Wilcoxon signed-rank
Association between variables
Chi-square
Causal relationships
Univariate regression analysis

Hypothesis Testing:1

Null Hypothesis (H0): The median number of hours of absenteeism taken by highly educated people is equal to the median number of hours of absenteeism taken by moderately educated people.

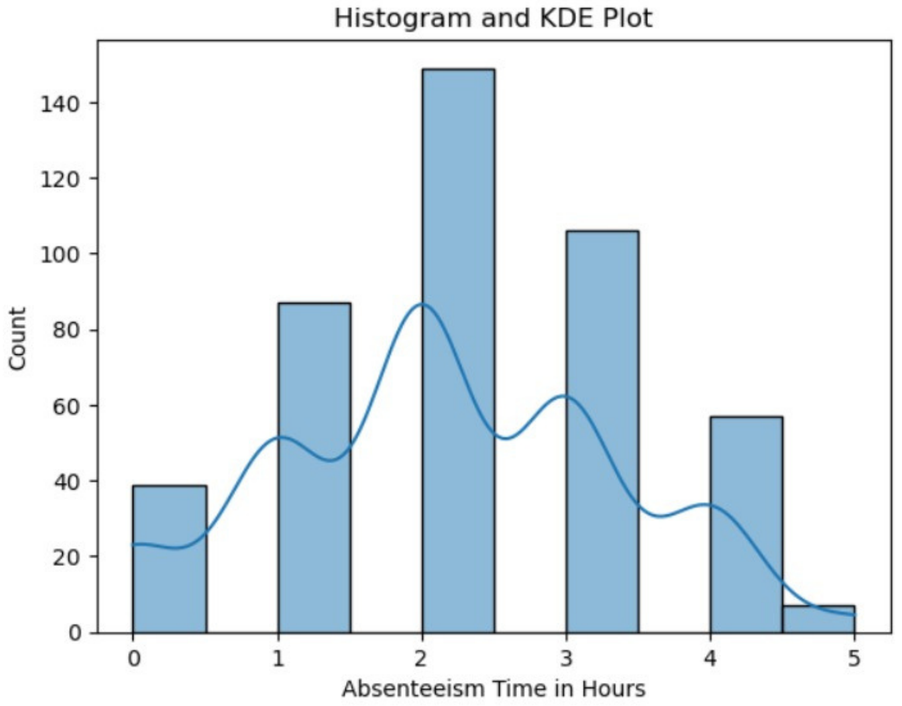
Alternative Hypothesis (H1): The median number of hours of absenteeism taken by highly educated people is not equal to the median number of hours of absenteeism taken by moderately educated people.

```python
import scipy.stats as stats
df_preprocessed=df_preprocessed[df_preprocessed["Absenteeism Time in Hours"]<=6]

import matplotlib.pyplot as plt
import seaborn as sns

data1=df_preprocessed["Absenteeism Time in Hours"]

from scipy.stats import shapiro


statistic, p_value = shapiro(data1)


print(f'Statistic: {statistic}, p-value: {p_value}')

alpha = 0.05
if p_value < alpha:
    print('The data does not follow a normal distribution (reject the null hypothesis).')
else:
    print('The data follows a normal distribution (fail to reject the null hypothesis).')
```

```
Statistic: 0.9305071830749512, p-value: 1.600080202883436e-13
The data does not follow a normal distribution (reject the null hypothesis).
```

Not a normal distribution-> non-parametric

```python
sns.histplot(data1, kde=True)
plt.title('Histogram and KDE Plot')
plt.show()
```



For independent/ unpaired samples - MannWhitney U Test

```
highly_educated = data2["Absenteeism Time in Hours"]
from scipy.stats import mannwhitneyu


statistic, p_value = mannwhitneyu(highly_educated, moderately_educated)

print(f'Mann-Whitney U Statistic: {statistic}')
print(f'P-value: {p_value}')


alpha = 0.05
if p_value < alpha:
    print('Reject the null hypothesis. There is a significant difference.')
else:
    print('Fail to reject the null hypothesis. There is no significant difference.')
```

```
Mann-Whitney U Statistic: 12702.0
P-value: 0.6588574282843959
Fail to reject the null hypothesis. There is no significant difference.
```

MODELING

Numerical Target Attribute

Absenteeism Time in Hours

CLASS 1 Boolean

Moderately absent

Excessively absent

CLASS 2 Boolean

Logistic Regression

- **Dependent Variable**

  absenteeism in working hours of employees

- **Independent variables**

  Reason 1, Reason 2, Reason 3,
  Reason 4, Month Value, Day of the Week, Transportation Expense, Distance to Work, Age, Daily Work Load Average, Body Mass Index, Education, Children, Pets

• Standardization

Accuracy Impact: Not standardizing the data before training results in a 10% decrease in accuracy.

Standardization Process: All features are standardized except:

- Four groups of the reason for absence.
- Two groups of the level of education.

Reasoning for Exclusion: The excluded variables are derived from dummy variables. Standardizing them may improve accuracy but compromises interpretability. The decision ensures a balance between accuracy and interpretability in the model.

• Interpreting the coefficients

	Feature name	Coefficient	Odds_ratio
3	Reason_3	2.952582	19.155348
1	Reason_1	2.618934	13.721092
2	Reason_2	0.834619	2.303937
4	Reason_4	0.644285	1.904625
7	Transportation Expense	0.621800	1.862277
13	Children	0.354918	1.426063
11	Body Mass Index	0.277050	1.319233
5	Month Value	0.011237	1.011300
10	Daily Work Load Average	-0.025833	0.974498
8	Distance to Work	-0.029342	0.971084
6	Day of the Week	-0.074809	0.927920
9	Age	-0.175852	0.838742
14	Pet	-0.274863	0.759676
12	Education	-0.293859	0.745382
0	Intercept	-1.431018	0.239065

Positive Correlation

- Reason_3
- Reason_1
- Reason_2
- Reason_4
- Transportation Expense
- Children
- BMI

Negative Correlation

- Age
- Pet
- Education

Neutral or No Impact:

- Month value :useful but not much
- Daily workload average
- Distance to work
- Day of the week

After Backward Elimation

	Feature name	Coefficient	Odds_ratio
3	Reason_3	2.940787	18.930743
1	Reason_1	2.602372	13.495716
2	Reason_2	0.843500	2.324489
4	Reason_4	0.637234	1.891243
6	Transportation Expense	0.619534	1.858062
10	Children	0.351950	1.421838
8	Body Mass Index	0.284103	1.328570
5	Month Value	0.005651	1.005667
7	Age	-0.176355	0.838320
9	Education	-0.263725	0.768185
11	Pet	-0.273698	0.760562
0	Intercept	-1.431381	0.238979

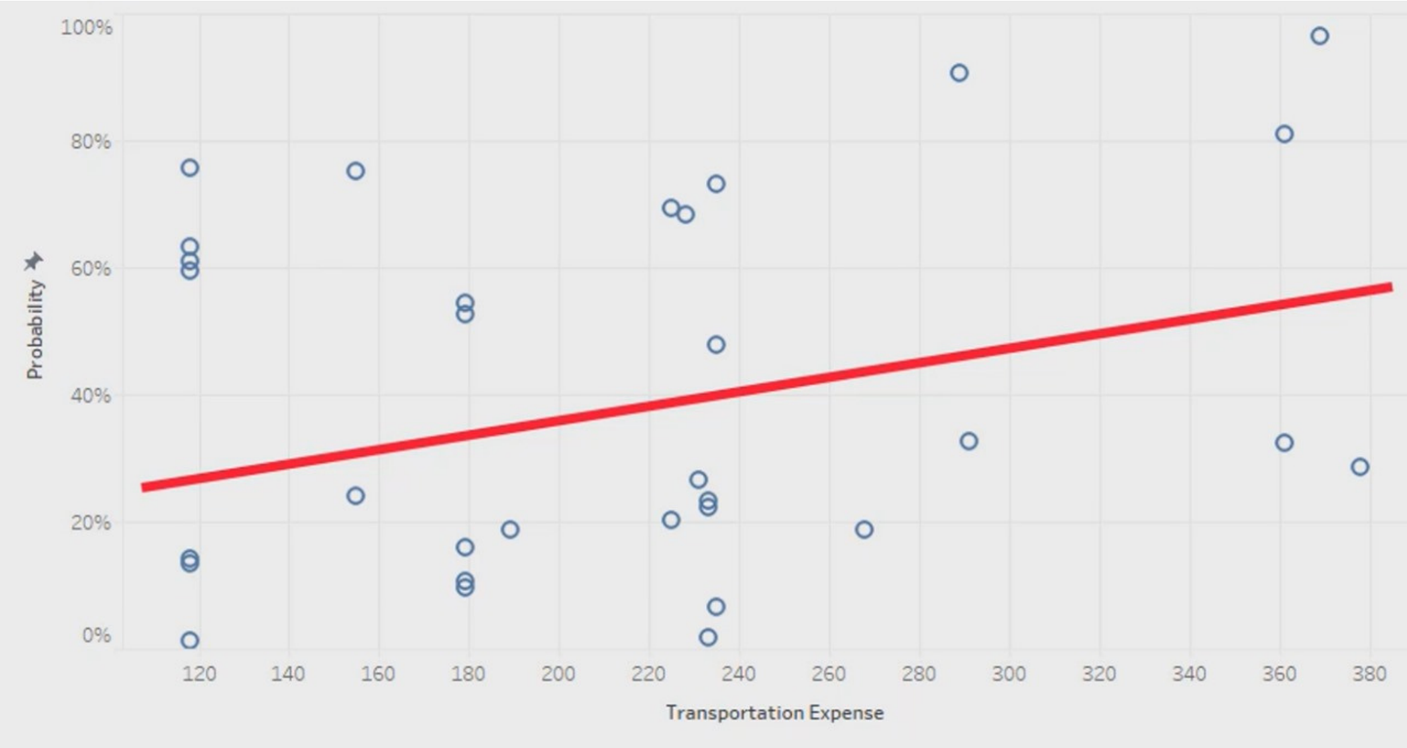
Reason 1: Various Diseases

Reason 2: pregnancy and giving birth

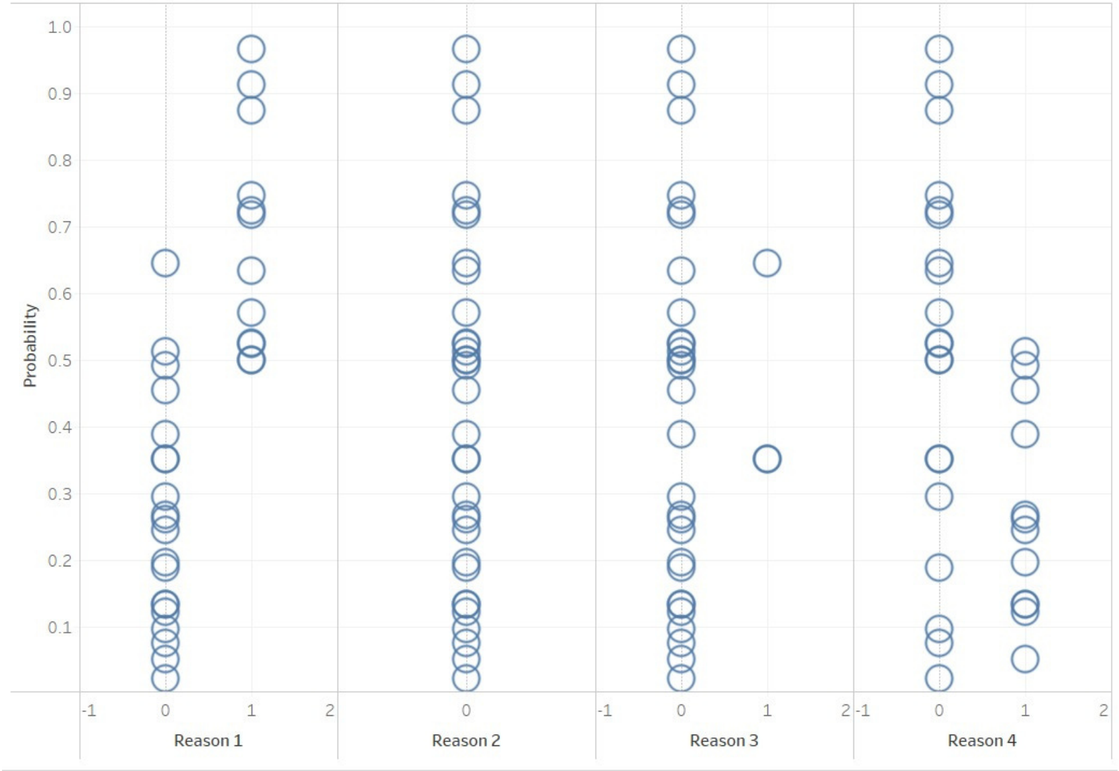
Reason 3: Poisoning

Reason 3: Light Disease

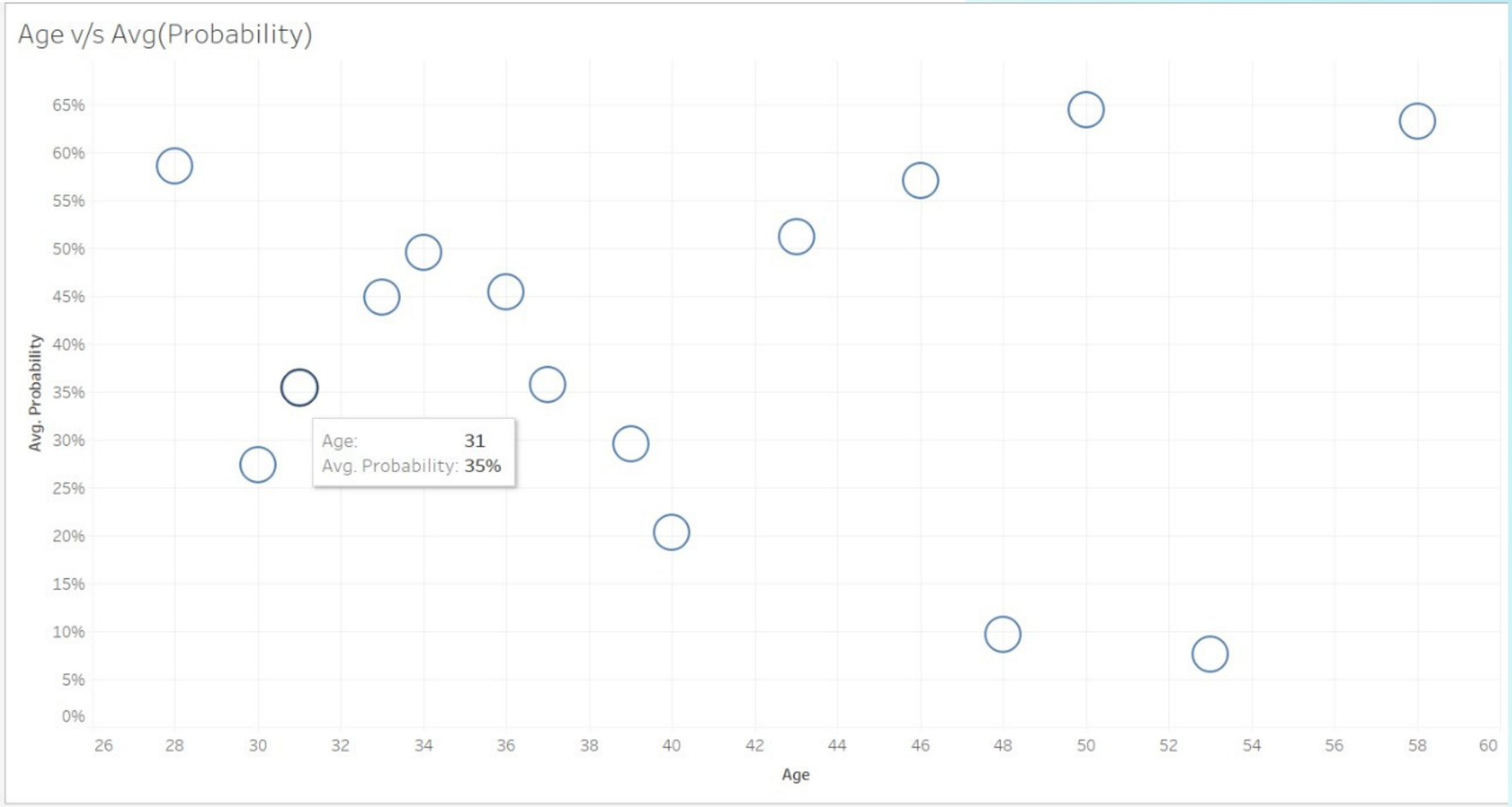
TRANSPORTATION EXPENSES AND CHILDREN



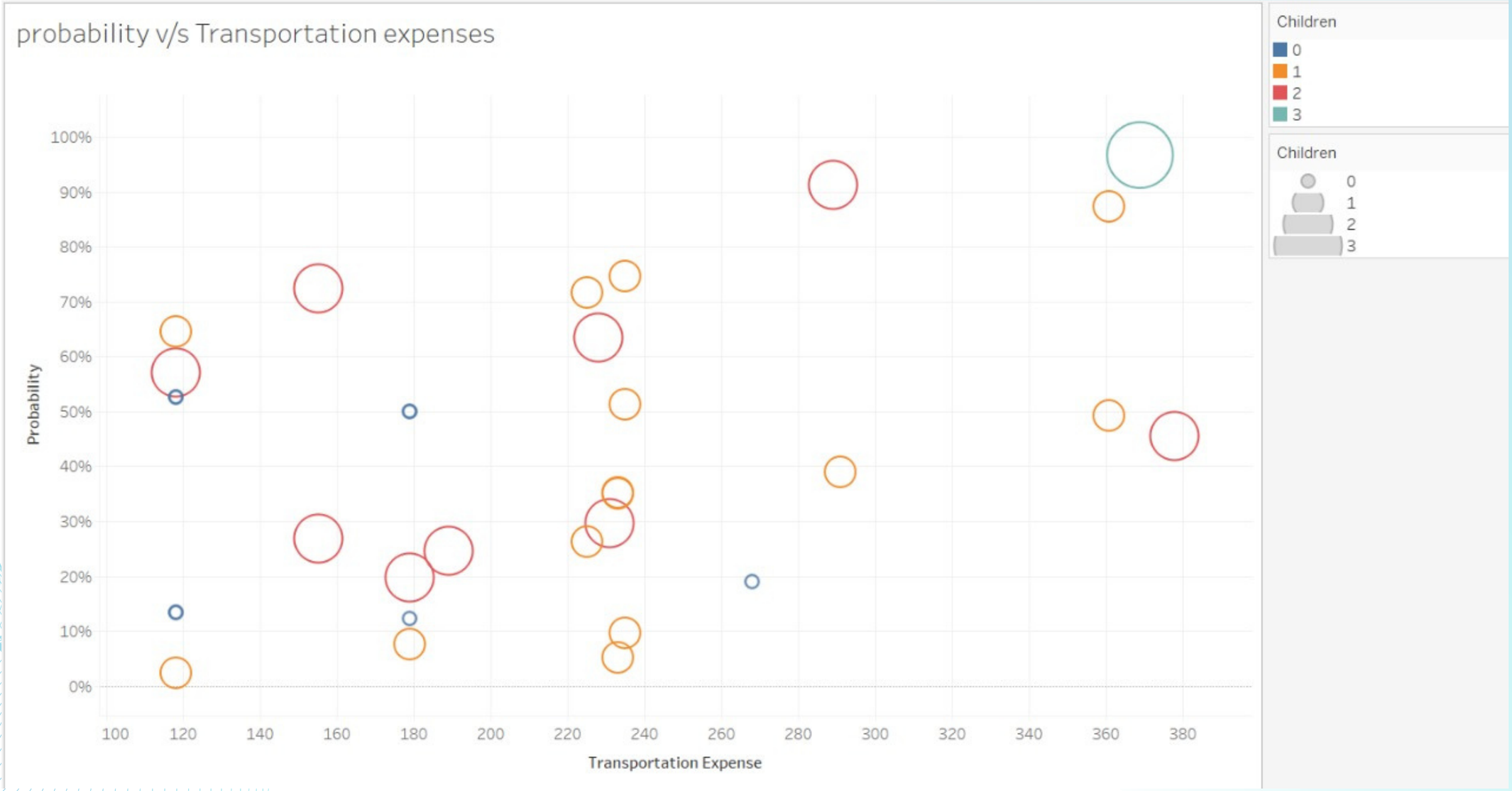
Probability v/s Four Reasons of Absence



TOPIC


Age v/s Avg(Probability)
Avg. Probability
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Age:
31
Avg. Probability:
35%
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
Age

TOPIC


probability v/s Transportation expenses
Probability
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
Transportation Expense
Children
0
1
2
3
Children
0
1
2
3

CONCLUSION

From an academic perspective, We have concluded that Data Analysis is a circulative process. As it is never a never-ending circle until we have reached a desired model. For example: the removal of outliers and Backward elimination.

For a productive manager,

- we can create special health programs for each age group to tackle their specific health issues.
- Provide transportation support or incentives.
- Introduce transportation subsidies or reimbursement programs
- On-Site Childcare and Pet Services

REFERENCES

- https://www.sciencedirect.com/science/article/pii/S0743731514000057
  Author:- Karthik Kambatla and Giorgos Kollias and Vipin Kumar and Ananth Grama

- Jupyter Notebook :- https://jupyter.org/
- Dataset:- https://drive.google.com/drive/my-drive

THANK YOU